```
                        ┌─────────────────────┐
                        │   Expression Data   │
                        └─────────────────────┘
          ┌───────────────┐    │
          ▼               │    │
┌─────────────────┐       │    │              ┌─────────────────┐
│ Class Discovery │───────────────────────────│  Known Classes  │
└─────────────────┘            │              └─────────────────┘
                               ▼
                    ┌─────────────────────┐
                    │ Assess Gene-Class   │
                    │    Correlation:     │
                    │ Neighborhood Analysis│
                    └─────────────────────┘
                               │
                               ▼
                    ┌─────────────────────┐
                    │   Build Predictor   │
                    └─────────────────────┘
                               │
                               ▼
                    ┌─────────────────────┐
                    │   Test Predictor    │
                    │ by Cross-Validation │
                    └─────────────────────┘
                               │
                               ▼
                    ┌─────────────────────┐
                    │   Test Predictor    │
                    │on Independent Dataset│
                    └─────────────────────┘
```

FIG. IA

$$c = (1,1,1,1,1,0,0,0,0,0,0)$$

$$gene_1 = (e_1, e_2, e_3, \ldots \ldots, e_{12})$$

$$gene_1 = (e_1, e_2, e_3, \ldots \ldots, e_{12})$$

C^*

C

FIG. 1B

|	AML	ALL	Weight
		V_1	W_1
		V_2	W_2
	V_3		W_3
		V_4	W_4
		V_5	W_5
	. . .	. . .	
	V_{AML}	V_{ALL}	
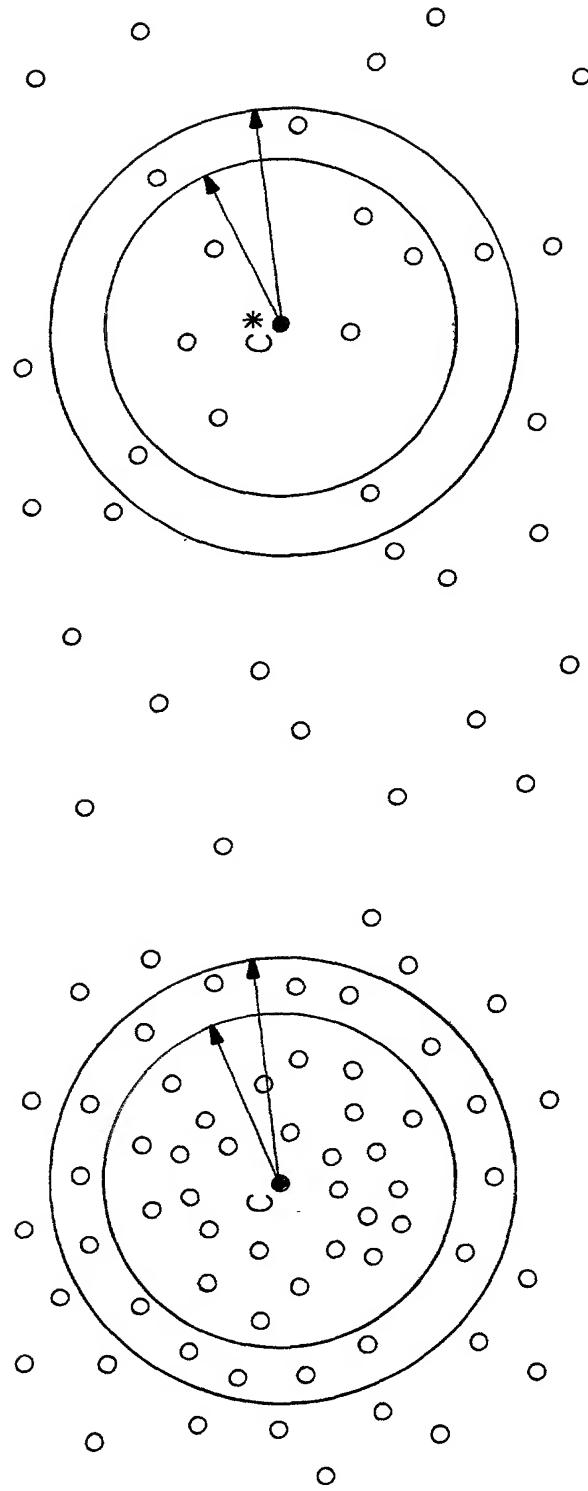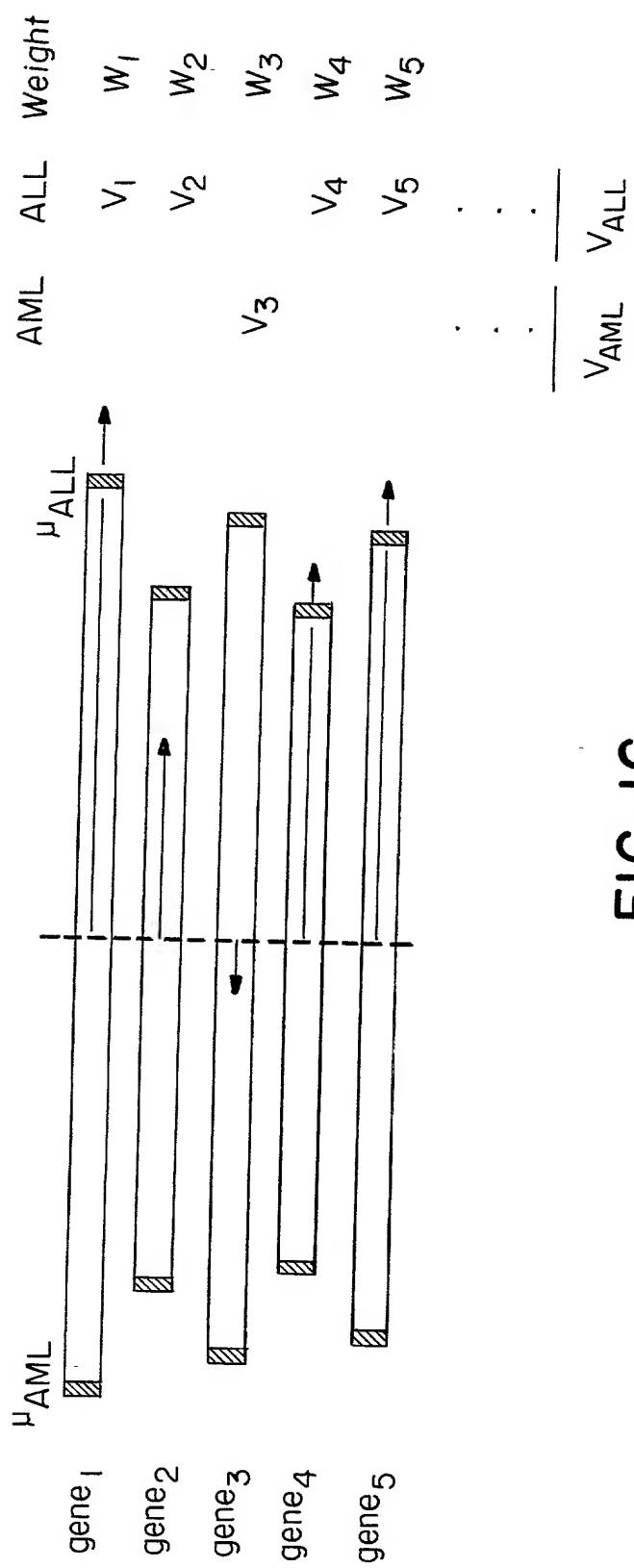
FIG. IC
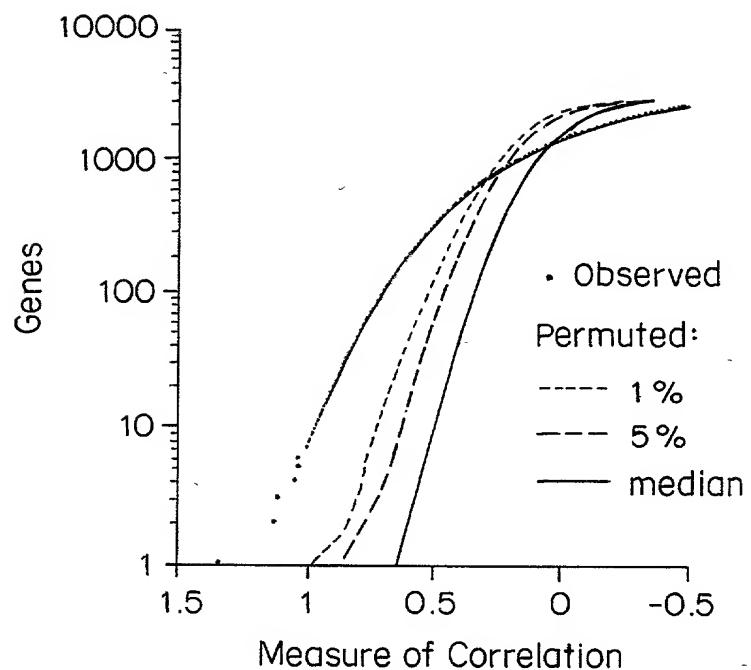
FIG. 2A
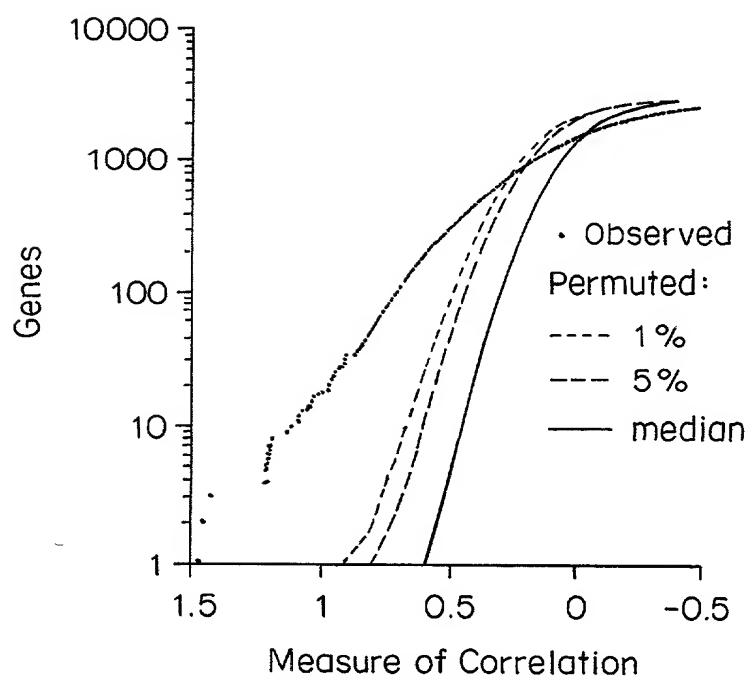
FIG. 2B
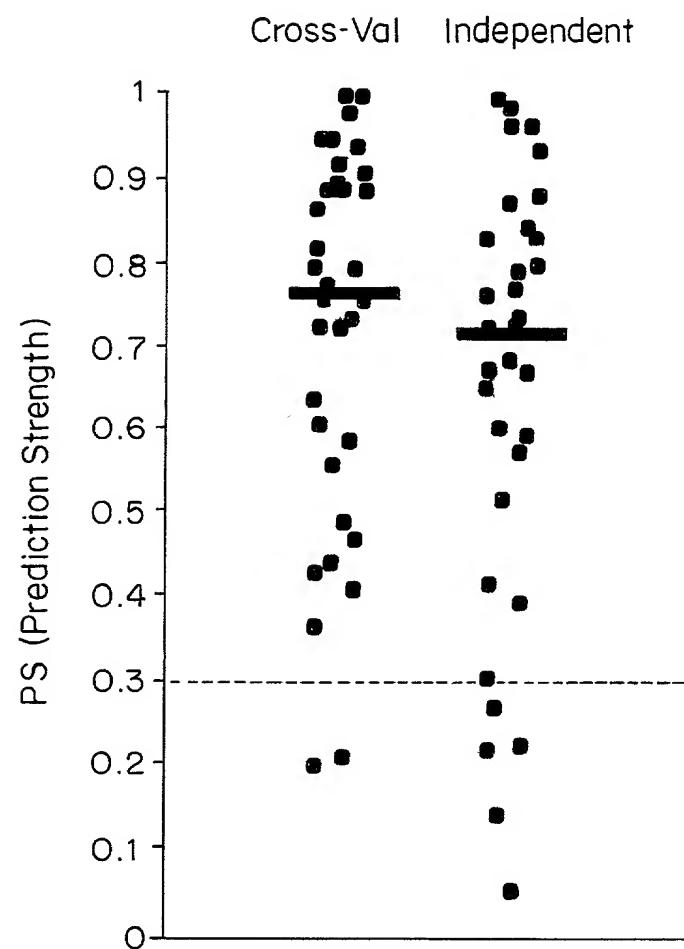
FIG. 3A
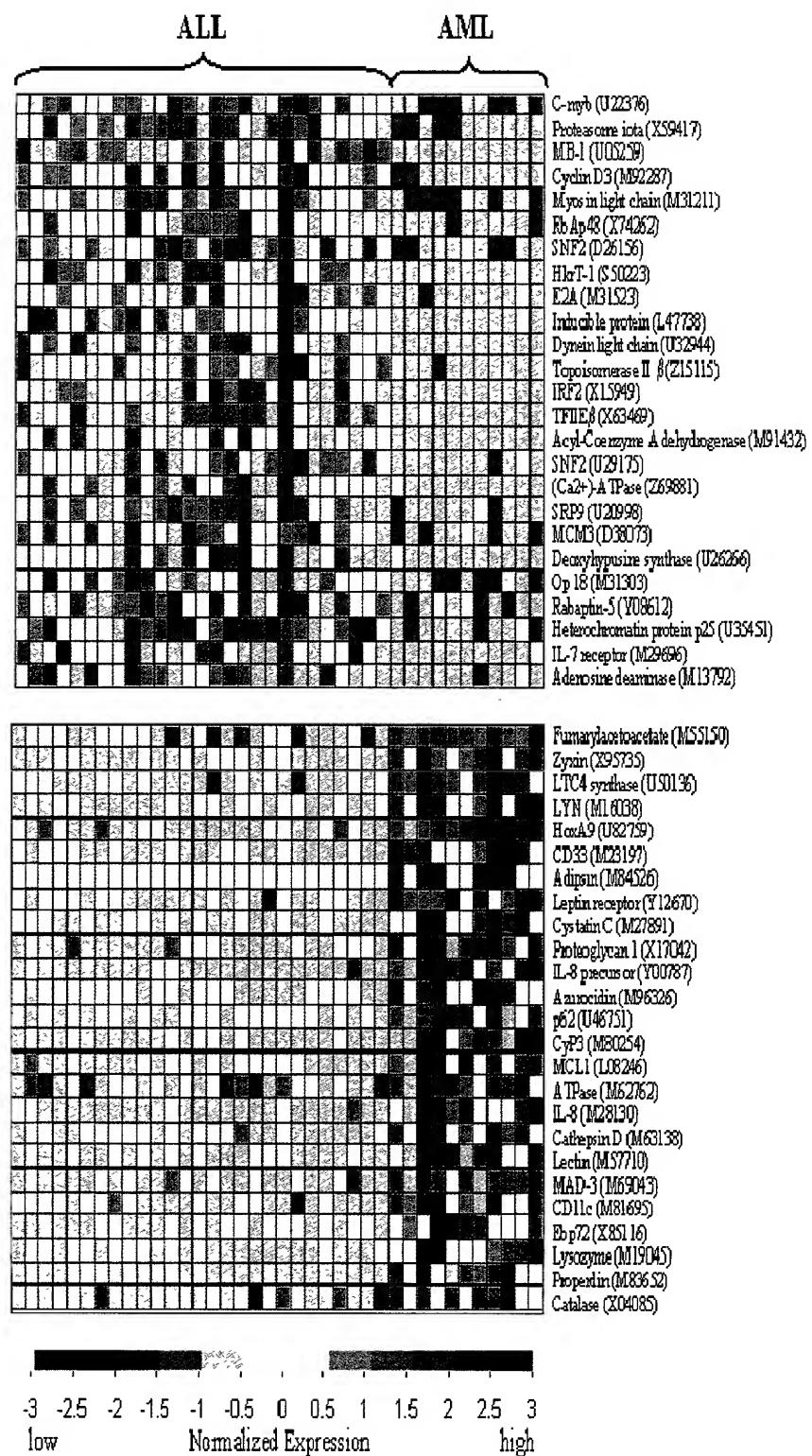
FIG. 3B
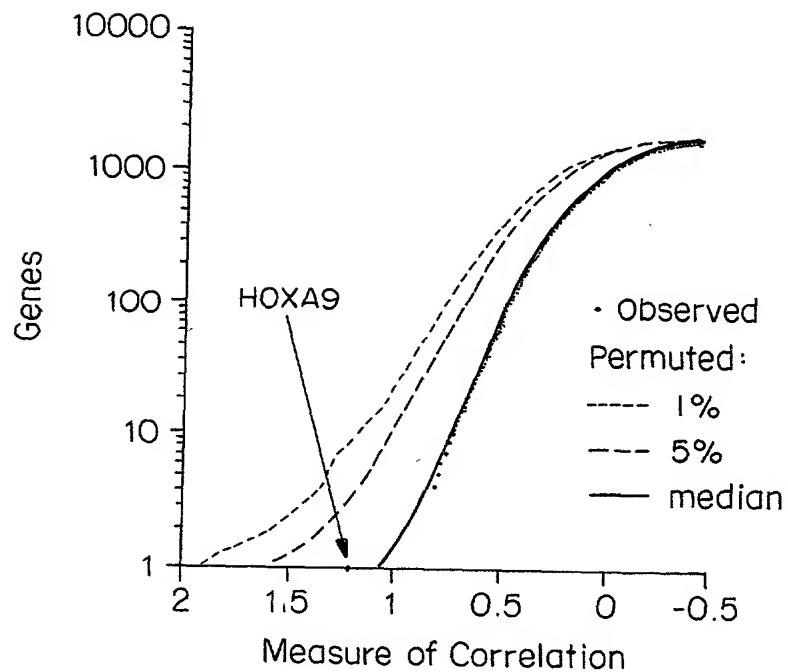
FIG. 4A
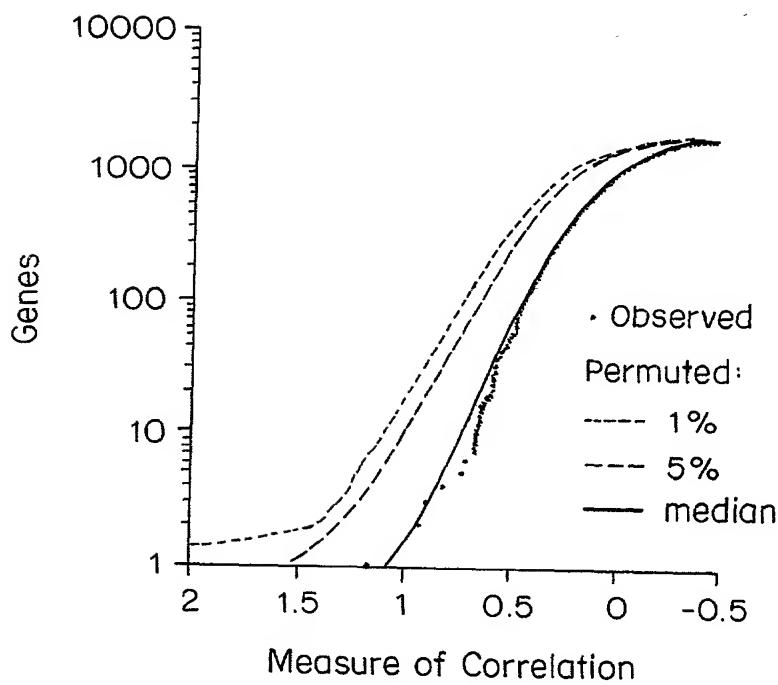
FIG. 4B
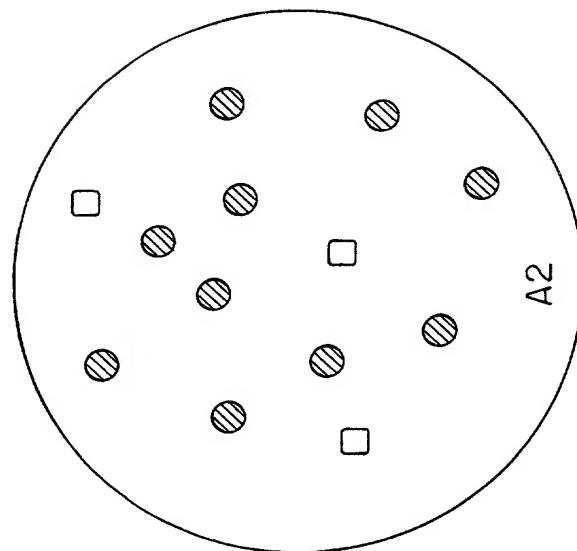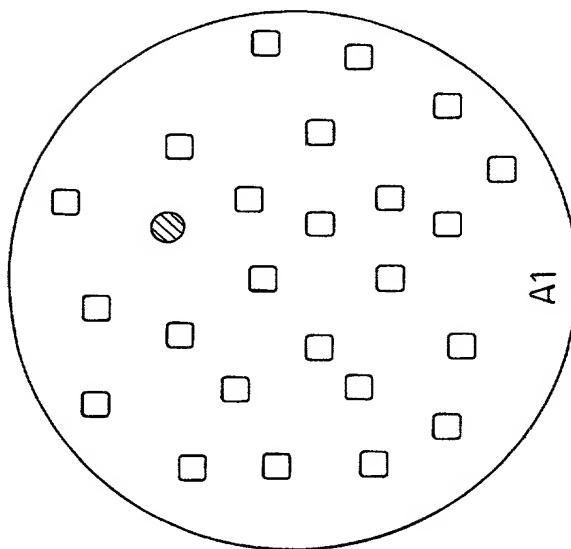
FIG. 5A
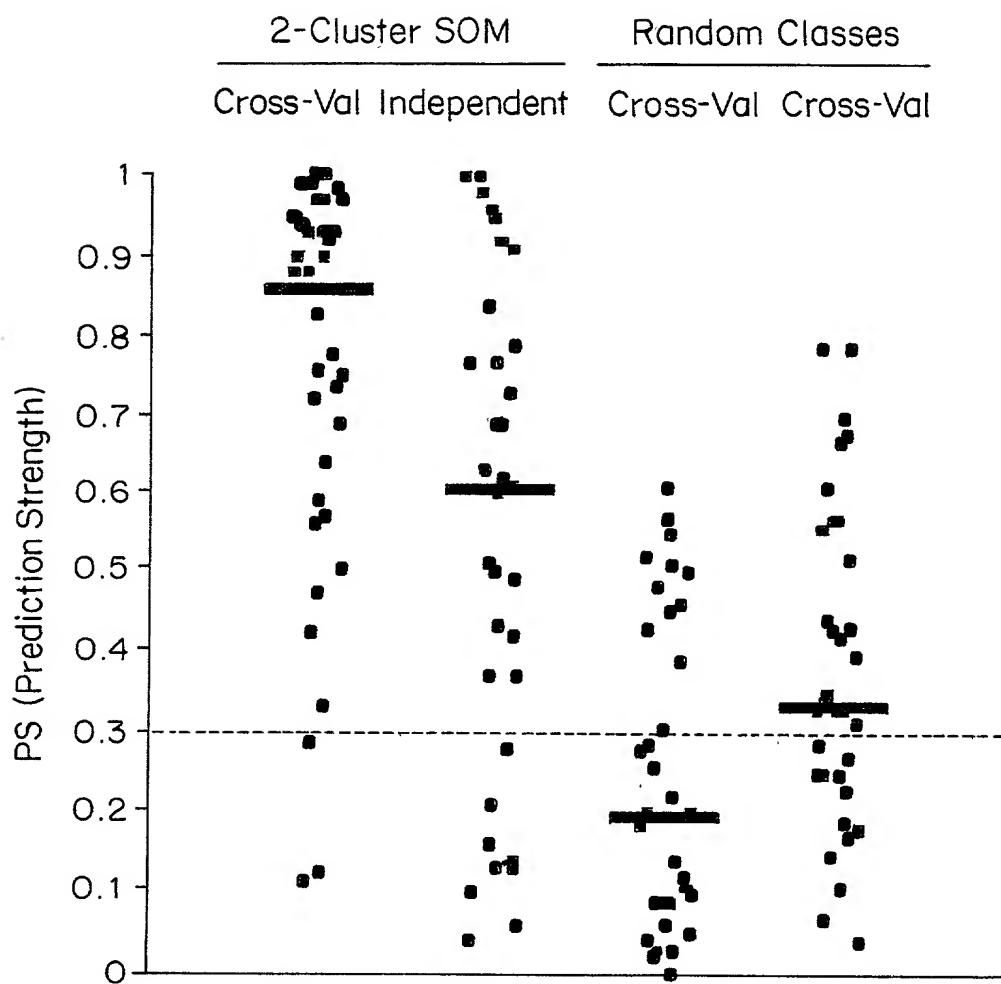
FIG. 5B
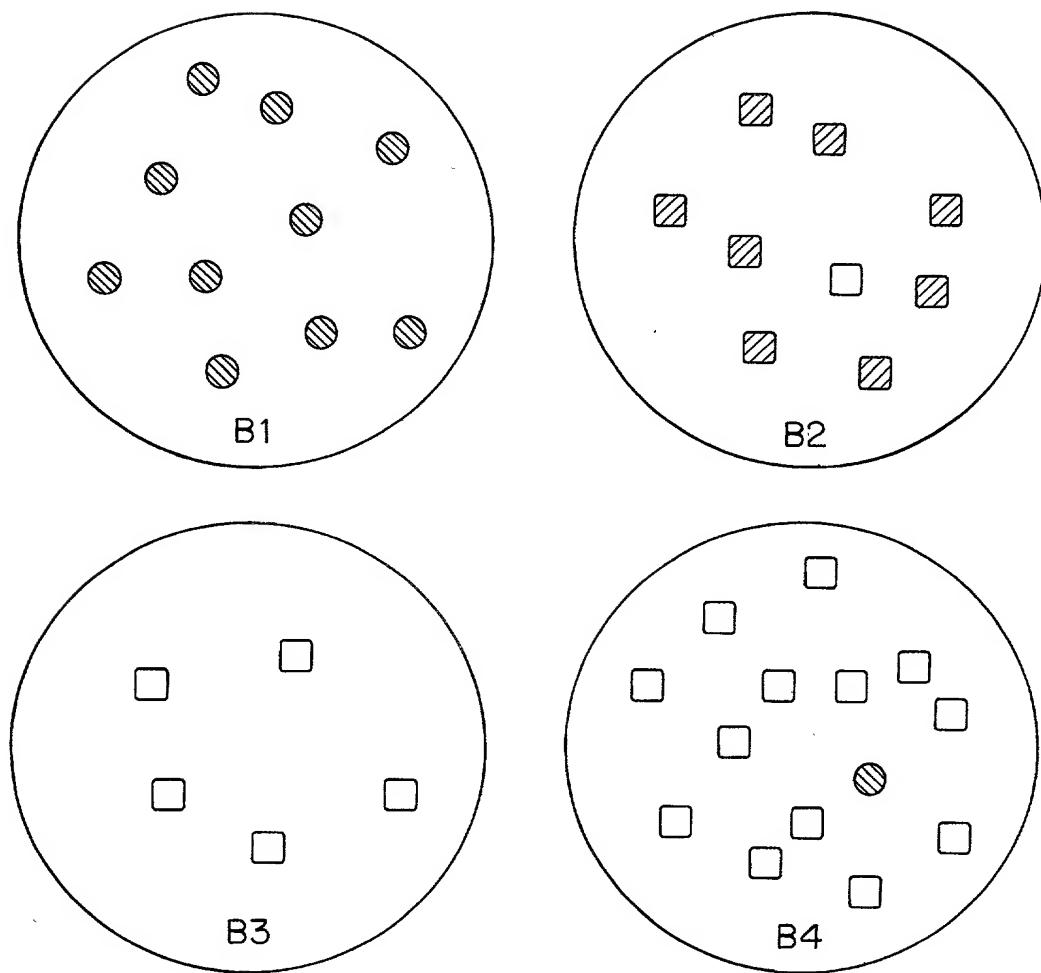
⊘ AML   ▨ T-ALL   ▢ B-ALL

FIG. 5C

FIG. 5D

Input Source 17

I/O Process
19

13

NN
15

Storage
21

I/O Process
23

Output
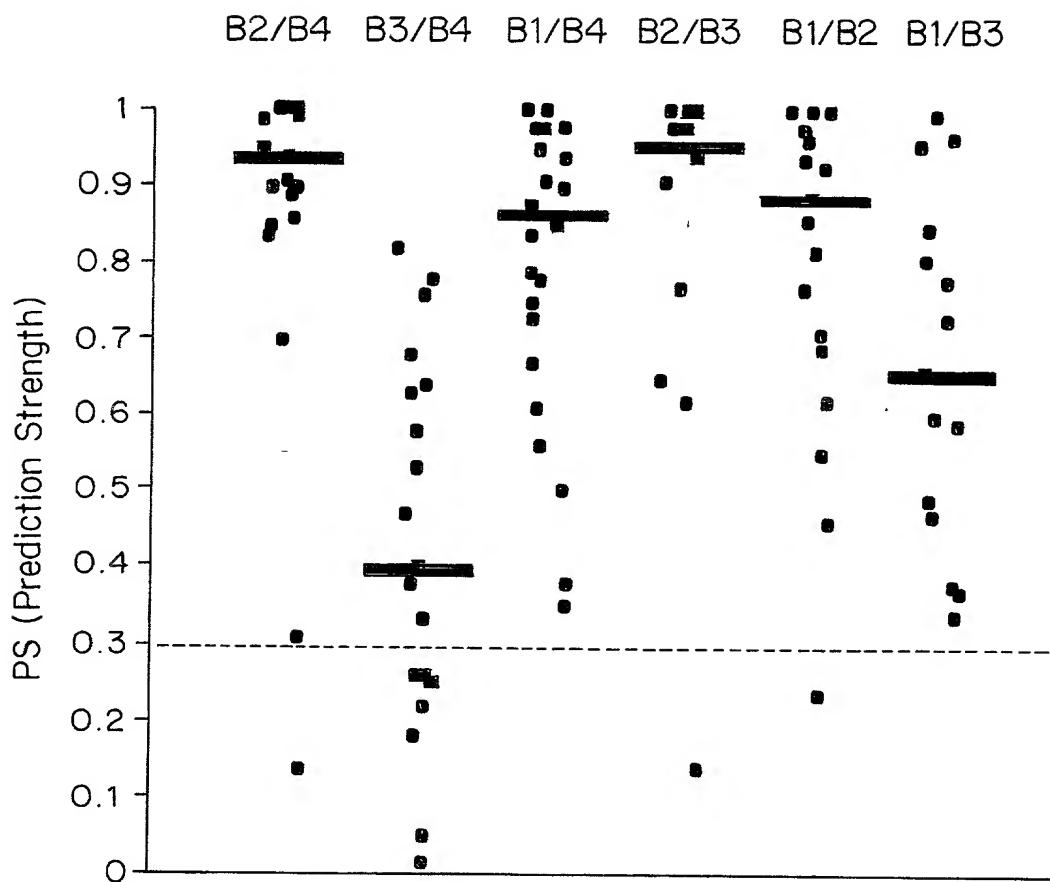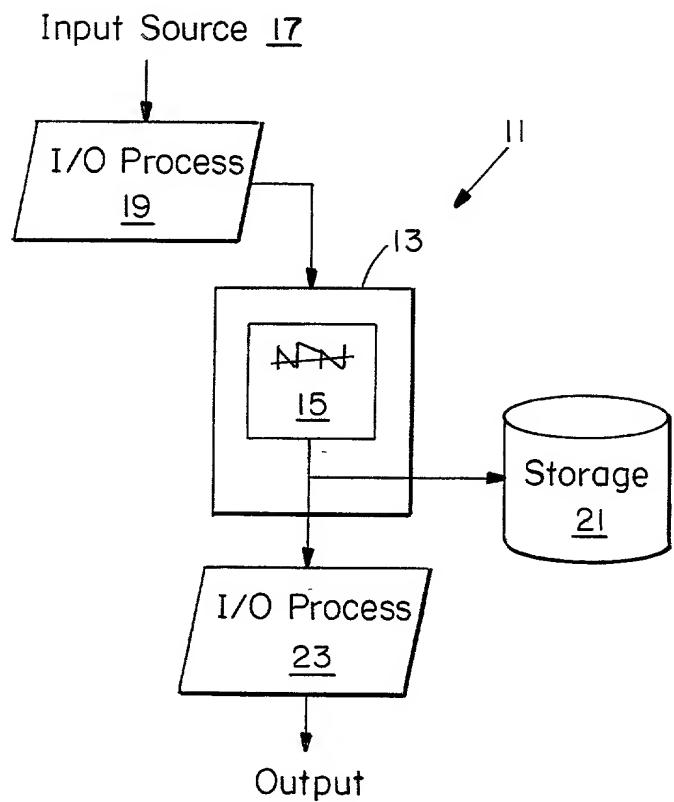
FIG. 6

Discovered Classes:
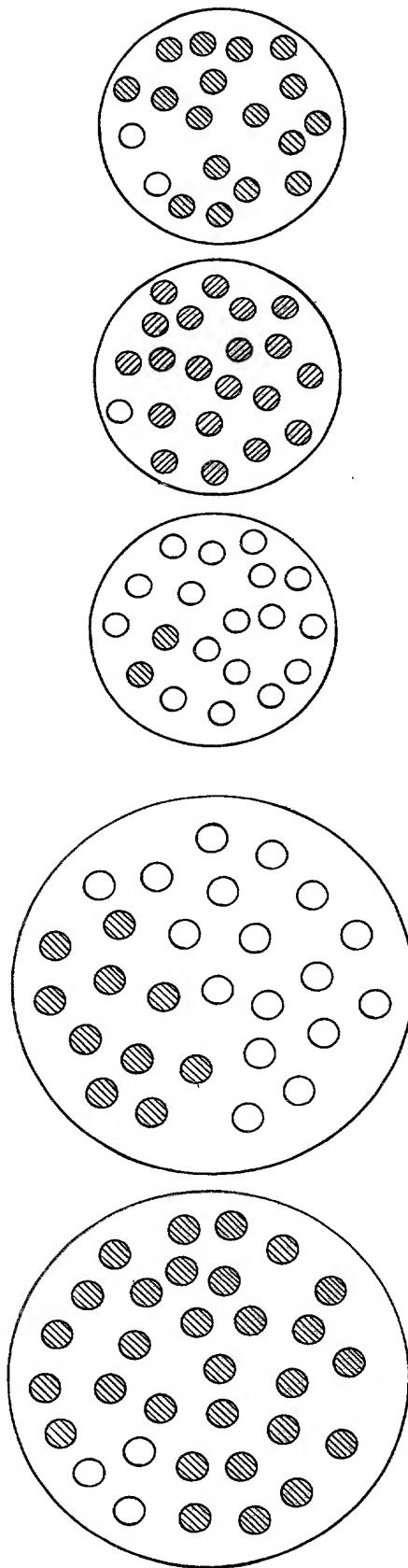
$\oslash, \circ$ Large B-Cell $\circ$ Follicular

$\oslash$ Large B-Cell

FIG. 7

Discovered Classes:

⊘ Medulloblastoma ◯ Glioma
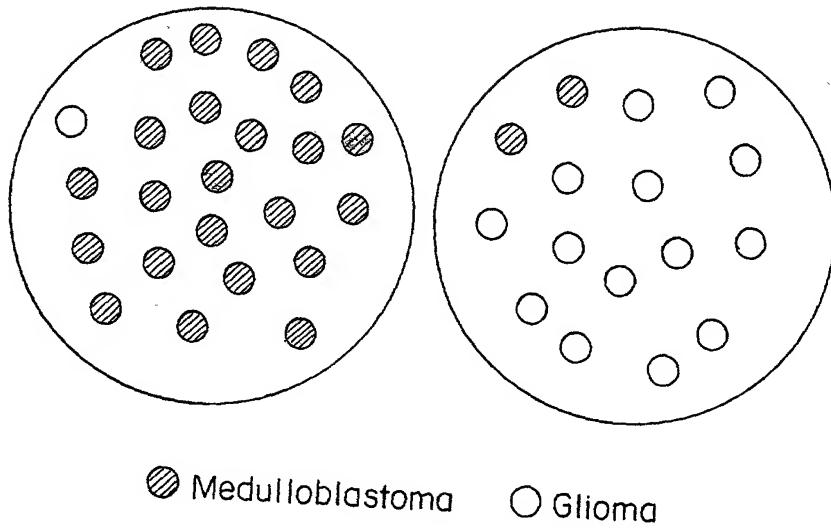
FIG. 8
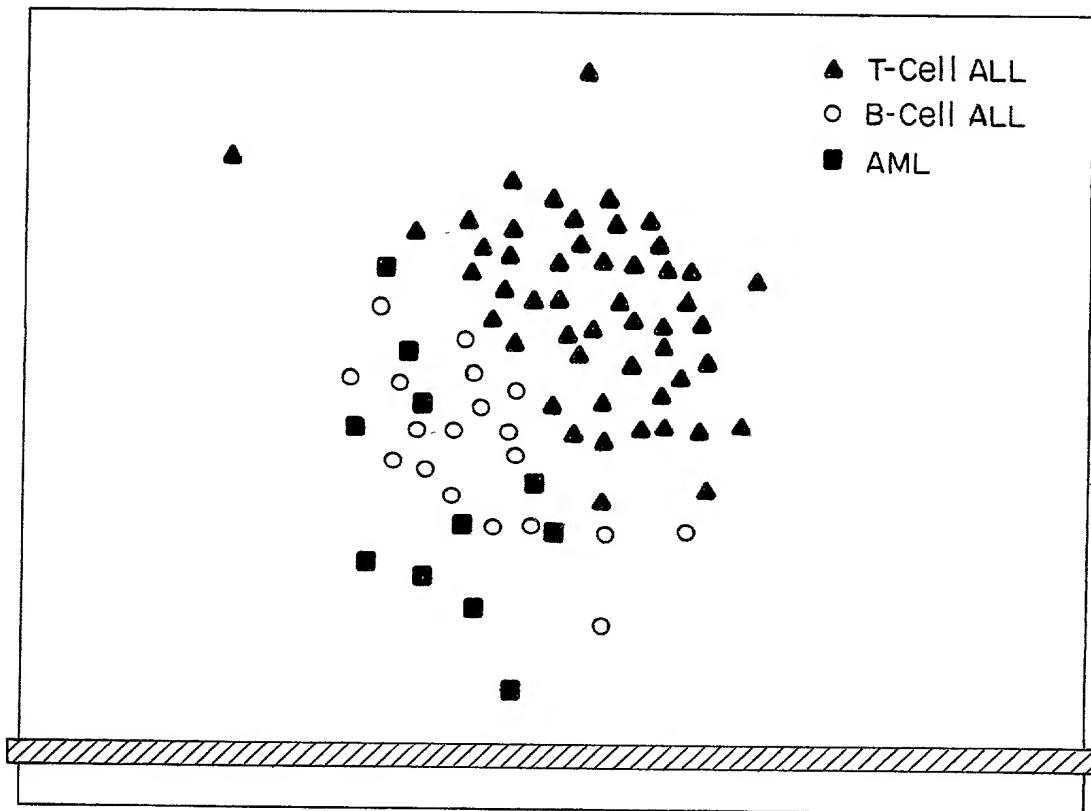
FIG. 9

FIG. 10

Hierarchy of Problems in Molecular Class Prediction

Problem:	Difficulty:	Gene Markers:	Error:	Example:
I. Tissue or Cell Type Normal vs. Abnormal	Low	~1000-2000	~0%	Normal vs. Renal Carcinoma
II. Morphological Type	Low-medium	~200-500	~0-5%	Leukemia ALL vs. AML
III. Morphological Subtype	Medium-high	~50-100	~0-15%	ALL B- vs. T-Cell
IV. Treatment Outcome Drug Sensitivity	High	~1-20	~5-50%	AML Treatment Outcome
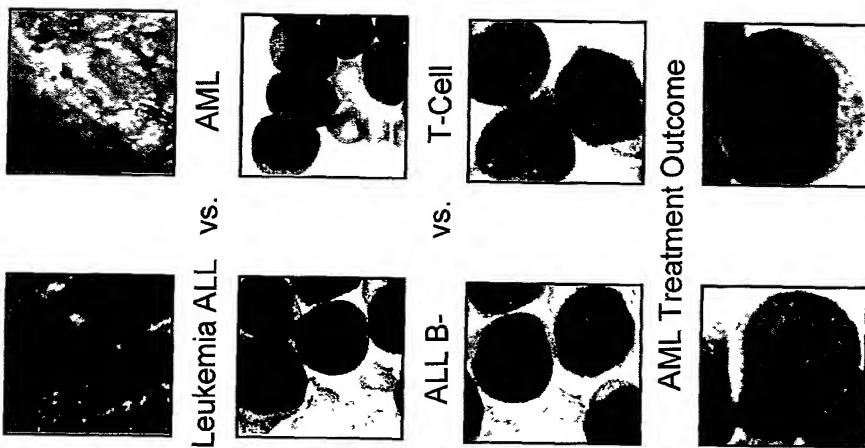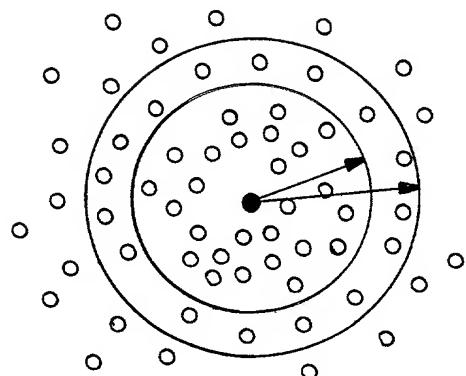
Class Pattern
Neighborhood

FIG. IIA
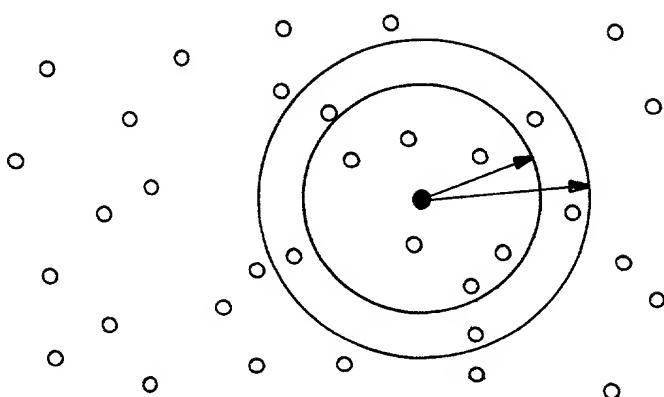
Permuted Pattern
Neighborhood

FIG. IIB

Freq.

$k = 1$

Density around
Permuted Pattern

Ref. 5% median 95%

Measure of Correlation

FIG. IIC

Ref. 5% median 95%

k

Significant
Neighbors
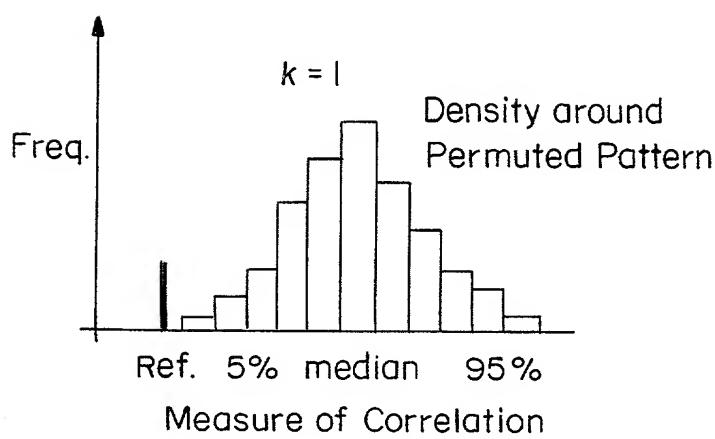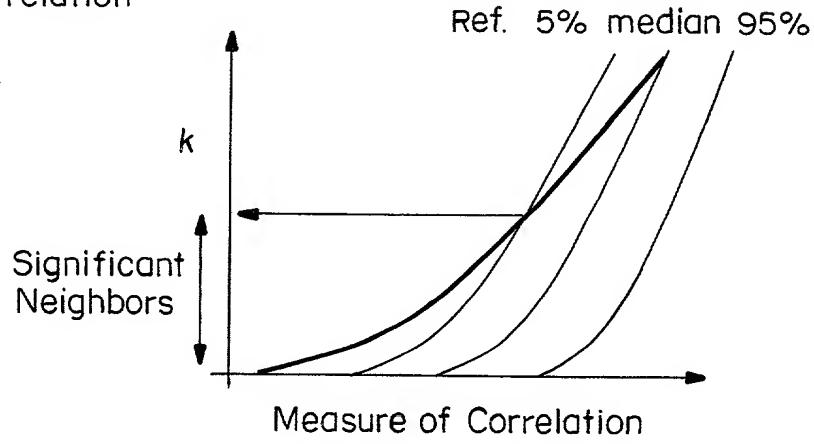
Measure of Correlation

## FIG. IID

CLASS PREDICTION RESULTS

Problem Type	Biological System	Problem Distinction	Number of samples	Number of errors	Number of no calls	Number of errors (all calls)	Number of gene markers
I	Renal	Normal v. Carcinoma	12	0 (0%)	0	0 (0%)	>100
II	Leukemia	ALL vs AML	35	0 (0%)	2	0 (0%)	700
III	Leukemia	ALL B v. T-Cell	33	0 (0%)	1	1 (3%)	200
IV	Leukemia	Treatment Outcome	15	2 (13%)	0	2 (13%)	1

Fig. 12